Data Wrangling with MongoDB
OpenStreetMap Project

Tianzhixi Yin

September 2015

1 Map Area

Lincoln, Nebraska, United States
https://www.openstreetmap.org/relation/169588

2 Problems Encountered in the Map

Street names

After downloading the OSM file for Lincoln, Nebraska, I first ran it against AuditStreetName.py to find all the irregular street types. I noticed some abbreviations like 'Ave', 'Dr', and 'St'. I updated them to the standard types, so 'W Kearney Ave' became 'W Kearney Avenue' and 'Pioneer Woods Dr' became 'Pioneer Woods Drive' and so on. This was done before I imported the data to MongoDB.

The following corrections were done in MongoDB:

1. people ignored the street types when they typed the street names. So I changed 'K', 'O', and 'P' to 'K Street', 'O Street', and 'P Street'. 'South 10th' was changed to 'South 10th Street', 'North 12th' was changed to 'North 12th Street', and 'NW 1st' was changed to 'Northwest 1st Street' (after making sure there are such street names in Lincoln).

2. people typed in a number after the street name. First I thought this was the house number mistyped here, but when I actually looked at all the records like this, house numbers were already there. I guess these are the room numbers that people don't know where to put them. Therefor, for instance 'NW 1st St 102' was changed to 'Northwest 1st Street' and I would suggest 'unit' be created with the value '102' (after making sure there is no original 'unit' input).

Country

I checked the 'country' part in addresses, all 60 inputs are correctly 'US'.

State

I checked the 'state' part in addresses, all 1069 inputs are correctly 'NE'.

City

I checked the 'city' part in addresses, the output is as follow:

```
{u'_id': u'Lincoln', u'count': 914}
{u'_id': u'Lincoln, NE', u'count': 3}
{u'_id': u'Seward', u'count': 1}
{u'_id': u'Denton', u'count': 1}
{u'_id': u'Malcolm', u'count': 1}
```

The only problematic entries are the three 'Lincoln, NE's, I changed all to 'Lincoln' and updated the 'state' part to 'NE' (I checked and found all three were missing the 'state' input).

Postcode

I sorted the postcodes by count , part of the output was as follow:

```
{u'_id': u'68508', u'count': 63}
{u'_id': u'68504', u'count': 53}
{u'_id': u'68462', u'count': 52}
......
{u'_id': u'68526', u'count': 2}
{u'_id': u'NE 68339', u'count': 1}
{u'_id': u'68523', u'count': 1}
{u'_id': u'68526-0736', u'count': 1}
{u'_id': u'68509', u'count': 1}
......
```

The only problematic one is 'NE 68339', I changed it to '68339' and updated the 'state' part to 'NE' (no state input originally).

House number

All are numbers, no obvious problem.

3 Data Overview

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

File sizes

lincoln_nebraska.osm..... 67 MB

lincoln_nebraska.osm.json .... 74 MB

Number of documents

```
> db.lincoln.find().count()
344308
```

Number of nodes

```
>  db.lincoln.find({"type":"node"}).count()
309717
```

Number of ways

```
> db.lincoln.find({"type":"way"}).count()
34589
```

Number of unique users

```
> len(db.lincoln.distinct("created.user"))
198
```

Top 1 contributing user

```
> result = db.lincoln.aggregate([{"$match":{"created.user":{"$exists":1}}},
                                 {"$group":{"_id":"$created.user","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":1}])
for ele in result:
    pprint.pprint(ele)

{u'_id': u'Your Village Maps', u'count': 148613}
```

Number of users appearing only once (having 1 post)

```
> result = db.lincoln.aggregate([{"$match":{"created.user":{"$exists":1}}},
                                 {"$group":{"_id":"$created.user","count":{"$sum":1}}},
{"$group":{"_id":"$count","num_users":{"$sum":1}}},
{"$sort":{"_id":1}},
{"$limit":1}])
for ele in result:
    pprint.pprint(ele)

{u'_id': 1, u'num_users': 41}
```

4 Additional Ideas About the Datasets

Problems in address information

As we can see in the auditing results, some sections are very clear for people to input information, like 'country', 'state', while others could cause inconsistency in the data. For example, people like to put the state name right behind the city name, if we specify in the 'city' section like 'no state name here, there is a state name section', maybe we can avoid this kind of inconsistency, save us the time for cleaning up this part. Similar situation for postcodes, if we put a message like 'no state name needed' or 'only the five numbers', this section would be more standardized.

So my main suggestion here is that the OpenStreetMap website can provide more detailed directions for people to input different information, or even have a small tutorial that shows the standards for data entries. This could save a great amount of time for auditing and correcting the data.

The main problem in addresses is the street name, if the OpenStreetMap website can display some good examples of street name entries for people (especially good examples for that particular area), I think contributors would tend to make less mistakes or nonstandard inputs. This might be a lot of work at the beginning since we have so many different areas in the world, but I believe it shall be worthy, for that can save us much more time in data wrangling.

For example, for the numbers after street names, my best guess is that those are the room numbers, but I could be wrong, if we really want to correct those, we should confirm with information from other sources. Such endeavours are very time consuming.

House name

When I checked the house names, part of the output was:

```
{u'_id': u'The Sawmill Building', u'count': 3}
{u'_id': u'Armour Building', u'count': 2}
{u'_id': u'Lincoln Hide & Fur Bldg', u'count': 2}
{u'_id': u'Village of Bennet', u'count': 1}
{u'_id': u'Suite #204', u'count': 1}
{u'_id': u'The Courtyards', u'count': 1}
......
```

The only seemingly error is 'Suite 204', after checking the record I realized maybe the person just mistyped the room number into 'housename'. However, I don't think it is necessary to change this one. So I would recommend renaming the 'housename' section to 'buildingname' because people might want to put the room numbers in 'housename' and that could be confused with 'suite' or 'unit'.

Suite

Only one record found, no problem present. The reason there is only one record here might be the existence of the 'unit' section, which is very similar to 'suite'.

Unit

No problem present, although some people seem to have confused 'unit' with 'suite', like the following example for 'unit':

```
{u'_id': u'D', u'count': 2}
{u'_id': u'Suite A', u'count': 1}
......
{u'_id': u'1,2', u'count': 1}
{u'_id': u'Suite J', u'count': 1}
......
```

But this is not a big problem.

Therefore, I would recommend replacing 'suite' and 'unit' with just 'room-number' because it seems like people are confused where to put the room number.

To sum up, I think the best way to reduce inconsistency is to provide more and better guidance for people who want to contribute to this project. A gamification idea could be rewarding users that make no or few entries that need to be updated or corrected.

Oftentimes the missing information could be found in other sections in the same node, if we prepare people well before they actually type in the information by showing them good or bad examples, I believe the cleaning work needed later on could be greatly reduced.

Contributor

The top 5 contributors have contributed 89.14% of the total data:

```
{u'_id': u'Your Village Maps', u'count': 148613}
{u'_id': u'behemoth14', u'count': 79833}
{u'_id': u'woodpeck_fixbot', u'count': 28289}
{u'_id': u'PHerison', u'count': 26407}
{u'_id': u'James GIS', u'count': 23786}
```

More than 40% of the users have only contributed 5 or less times. If we can find ways to encourage ordinary people to contribute, the map should be more detailed and useful. Another idea I have is to let other contributors update or correct the existing data. There might be more than one person who are familiar with a particular area and they are the best candidates to seek out the errors or outdated information within the map. And also, we can rewards people that make a lot of good corrections.

Potential issues of my suggestions

One potential problem could be that people from different parts of the world do not input or read information the same way so it is hard to come up with an overall perfect guidance for the whole website. However, I still think good guidance is necessary so maybe we can first develop guidance for big areas such as a country or for a particular language.

Another potential problem I can think of is in the gamification system. Sometimes people would try any sorts of things to achieve more badges or higher place in the leaderboard. So maybe there will be people who put in worthless information just to add to their seemingly achievements or even write programs to contribute, which is maybe not a good thing for now. So besides the gamification system, maybe a punishment system is also needed to prevent noise in the data.

Finally, for my idea of letting peers review and update the data, maybe there will be times when people disagree. If there is a controversy or argument about certain entries, people should be able to provide outside information, cross-reference or maybe even let other users have a majority vote to decide. If most local people agree on one issue then that should be the standard. Although such scenarios might not happen often but as we solve more and more difficulties and set up a better standard for everyone, the OpenStreetMap project shall be more valuable and attract more people to contribute.

Additional data exploration using MongoDB queries

Top 5 appearing amenities

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}}},
                                 {"$group":{"_id":"$amenity","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)
```

```
{u'_id': u'parking', u'count': 567}
{u'_id': u'place_of_worship', u'count': 228}
{u'_id': u'school', u'count': 138}
{u'_id': u'fast_food', u'count': 137}
{u'_id': u'restaurant', u'count': 104}
```

Top 5 religions

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"place_of_wo
                                 {"$group":{"_id":"$religion","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
```

```
    pprint.pprint(ele)

{u'_id': u'christian', u'count': 218}
{u'_id': None, u'count': 4}
{u'_id': u'buddhist', u'count': 3}
{u'_id': u'jewish', u'count': 1}
{u'_id': u'muslim', u'count': 1}
```

Top 5 denominations

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"place_of_wor
                                {"$group":{"_id":"$denomination","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)

{u'_id': None, u'count': 119}
{u'_id': u'lutheran', u'count': 30}
{u'_id': u'methodist', u'count': 23}
{u'_id': u'baptist', u'count': 18}
{u'_id': u'catholic', u'count': 13}
```

Maybe many churches are non-denominational.

Top 5 cuisines in restaurants

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"restaurant"}
                                {"$group":{"_id":"$cuisine","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)

{u'_id': u'american', u'count': 27}
{u'_id': None, u'count': 15}
{u'_id': u'pizza', u'count': 8}
{u'_id': u'chinese', u'count': 8}
{u'_id': u'italian', u'count': 5}
```

Top 5 cuisines in fast foods

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"fast_food"}}
                                {"$group":{"_id":"$cuisine","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)
```

```
{u'_id': u'sandwich', u'count': 30}
{u'_id': u'burger', u'count': 24}
{u'_id': u'ice_cream', u'count': 16}
{u'_id': u'mexican', u'count': 15}
{u'_id': None, u'count': 14}
```

No surprise here.

5 Conclusion

I would say the data in Lincoln, NE are pretty well documented, not a lot of errors were found. The people who contributed to this area did a great job. However, if we want this map to be perfect, there are still some work to be done, especially some cleanings need extra information outside of OpenStreetMap to be made. Also, the data format could be improved a little by replacing the sections that are confusing for people. Overall, this is a fun project and I truly learned a lot.